Les bases de données « Ressources génétiques » sur Internet : possibilités d'accès, contraintes actuelles et perspectives

Roland Cottin, Emmanuelle Alfonsi, Dominique Agostini

La nécessité d'identifier et de répertorier les ressources génétiques au niveau mondial, dans un but de conservation et d'enrichissement de la biodiversité, n'est plus à démontrer. Les nombreuses recherches sur la description des ressources génétiques, qu'elles soient d'origine animale, végétale ou microbienne, génèrent une masse importante d'informations, de plus en plus gérées sous la forme de bases de données informatisées. La mise à disposition publique de la connaissance du patrimoine génétique peut se faire aujourd'hui *via* Internet qui est un support privilégié d'échange de l'information scientifique et technique. Dès lors que l'on associe un domaine comme les ressources génétiques à un système informatique support et gestionnaire de cette information et à un vecteur de communication comme Internet, on dispose d'un outil aux potentialités énormes. Cependant, ce monde évolue tellement rapidement qu'il est apparu nécessaire de faire un état des lieux des supports d'information existants, d'identifier les outils permettant d'y accéder, et d'analyser le type d'information consultable dans les bases de données de ressources génétiques.

Après une présentation des différents supports de l'information qui constituent Internet ainsi que celles des moteurs de recherche utilisables avec leurs limites, cet article expose les caractéristiques principales des bases de données sur les ressources génétiques actuellement sur Internet et les évolutions en cours dans ce domaine.

R. Cottin, D. Agostini : SRA INRA-CIRAD, 20230 San Giuliano, France.
E. Alfonsi : Unité régionale de documentation, Centre INRA de Corse, 20230 San Giuliano, France.

Tirés à part : R. Cottin

Matériels et méthodes

Internet, réseau de réseaux, est constitué d'un ensemble d'ordinateurs [1] concrètement reliés entre eux selon l'image de la toile d'araignée mondiale *(World Wide Web)* dont www est l'acronyme. Ces machines communiquent entre elles, de façon plus ou moins invisible pour l'utilisateur, selon de nombreux protocoles qui sont autant de langages différents. Rechercher de l'information sur Internet oblige à prendre en compte cette diversité pour atteindre le résultat attendu. La médiatisation et l'accès de tous à Internet, créé à l'origine pour les militaires américains en 1969 sous le nom d'Arpanet *(Advanced Research Project Administration NETwork)*, utilisé ensuite par les scientifiques, ont rendu obsolètes des outils simples et d'aspect rustique [2], fonctionnant principalement en mode texte, au profit d'interfaces graphiques plus aisées à utiliser. Néanmoins, il reste encore de nombreuses ressources uniquement accessibles sous ces anciens formats.

Moteurs de recherche interrogés

Internet étant non structuré par construction, la recherche d'une information peut rapidement aboutir, sans guide pour l'utilisateur, à une submersion complète. Plusieurs outils ont été développés [3] afin d'aider l'internaute à rechercher des informations par l'utilisation de mots-clés. Cependant, une même requête, soumise à deux moteurs de recherche différents, ne donnera pas forcément le même résultat. Le moyen le plus simple pour trouver une information sur Internet est d'utiliser un serveur spécialisé dans la localisation de ressources sur le Web : le moteur de recherche. Il en existe plusieurs dizaines, certains à vocation généraliste, d'autres plus spécialisés soit thématiquement, soit géographiquement.

Le plus souvent, les résultats de recherche obtenus par ces outils ne sont pas directement exploitables par l'utilisateur. En effet, le résultat peut être constitué, d'un extrême à l'autre, d'une liste vide ou alors de plusieurs milliers de références classées dans l'ordre de leur découverte. De plus, la pertinence de la réponse ne dépend pas uniquement de la précision de la demande réalisée : par exemple une recherche sur le mot « Citrus », entraîne une avalanche de références sur les annonces des agences immobilières d'un comté de Floride portant le même nom…

Des logiciels spécialisés sont apparus sur le marché : les agents de recherche. Ils

Encadré 1

Analyse des différents systèmes supports de l'information

Vouloir réaliser l'inventaire des bases de données concernant les ressources génétiques sur Internet nous a conduit, en premier lieu, à analyser les différents supports d'information existants ainsi que les outils de communication qui leur sont respectivement associés. Ces ensembles « supports-outils » sont présentés par ordre croissant de facilité d'utilisation et d'accès à une information élaborée.

1. Telnet
Ce moyen de communication basique permet de se connecter en tant que console, sur un ordinateur distant. Cela nécessite de disposer obligatoirement d'un nom d'utilisateur et d'un mot de passe valide. Une fois la connexion réalisée, l'utilisation des commandes systèmes et des programmes disponibles sur l'ordinateur distant permet d'accéder aux données recherchées. Le caractère un peu archaïque de l'interface et les problèmes de sécurité liés à une connexion sur un ordinateur distant limitent ce moyen de communication à l'échange d'informations de façon interne au sein d'un organisme, sans ouverture vers le public.

2. Gopher
Le Gopher est un système développé au début des années 90, à l'Université du Minnesota (États-Unis). Son nom provient d'ailleurs de la mascotte de cette dernière (*gopher*, écureuil en français). Gopher a été conçu comme un outil de recherche pour faciliter la navigation au sein d'arborescences, représentant l'organisation des niveaux de l'information, indépendamment de leur origine géographique. Bien qu'accessible par interface textuelle *(figure 1)*, il ne nécessite pas l'utilisation de commandes complexes, puisque seules les flèches de déplacement et la touche « Entrée » sont nécessaires pour naviguer dans l'arborescence et consulter l'information. Avec le développement du WWW qui permet d'accéder au même type d'information sous une interface graphique, les serveurs basés sur Gopher sont actuellement en perte de vitesse ; et ceci malgré le développement récent d'un outil dénommé « Veronica » permettant l'utilisation des opérateurs booléens (AND, OR et NOT) pour affiner la recherche et la pertinence de l'information récupérée.

3. FTP
Le FTP *(File Transfer Protocol)* est un outil permettant de recevoir et d'envoyer des données entre ordinateurs reliés par Internet. Généralement, et contrairement à Telnet, des accès en « anonyme » sont ouverts sur les serveurs. Ils ne nécessitent pas l'attribution d'un nom d'utilisateur et d'un mot de passe à chaque visiteur potentiel. Un logiciel spécialisé et une connaissance des commandes de base sont nécessaires pour accéder aux fichiers disponibles sur les serveurs *(figure 2)*. Ce protocole présente l'avantage de pouvoir transférer des fichiers de taille illimitée, contrairement au e-mail (courrier électronique), par exemple. Comme pour le Gopher, l'interface WWW intègre de plus en plus le protocole FTP et facilite son utilisation en évitant le recours à un logiciel spécialisé.

4. *Mailing-list* ou liste de diffusion
La *Mailing-list*, ou liste de diffusion, définit un espace de communication accessible par l'utilisation des messages e-Mail. Concrètement, il suffit de s'abonner, généralement gratuitement, à une *mailing-list* en envoyant simplement le mot *« subscribe »* à l'adresse du serveur *(tableau 1)* qui en est gestionnaire pour recevoir tous les messages envoyés par les autres abonnés. Une liste de diffusion peut être libre ou modérée. Elle est dite modérée lorsque le propriétaire de la liste valide chaque message avant qu'il ne soit diffusé aux membres. Certaines listes peuvent être privées et réservées à un groupe d'utilisateurs, cependant la plupart d'entre elles sont publiques et donc ouvertes à tous. Ce vecteur d'information peut être très riche, mais il peut aussi générer un « bruit de fond » important. C'est un lieu idéal pour déclencher des polémiques qui entraînent alors une avalanche de réactions, dont le principal inconvénient sera d'encombrer les boîtes aux lettres de la totalité des abonnés.

5. *Newsgroups* ou forum de discussion
Proche de la *mailing-list*, le *newsgroup* ou forum de discussion ne se distingue principalement de cette dernière que par le protocole de communication utilisé. L'information est disponible sur des serveurs dédiés à cet usage (les serveurs de news) et son accès nécessite un logiciel spécialisé *(figure 3)*. Les thèmes les plus divers sont abordés au sein de ses *newsgroups*. Théoriquement, n'importe qui peut créer son propre forum de discussion. Ceci se traduit par un nombre impressionnant de thèmes, et certains serveurs affichent plus de 15 000 forums de discussion en ligne. Cependant, pour que la diffusion d'un thème soit généralisée, cela nécessite qu'elle soit relayée par d'autres serveurs de *news*. Pour se retrouver dans la profusion des forums, leur dénomination doit obéir à des règles précises, identifiant de façon résumée le thème traité. Les principaux forums afférents aux ressources génétiques sont indiqués dans le *tableau 2*.

6. Web
C'est sans doute le système, associant une interface et un réseau, le plus connu du public et souvent assimilé de façon abusive à la totalité d'Internet *(figure 4)*. Historiquement, la première interface du Web (l'explorateur) est redevable aux travaux de Tim Berners-Lee, informaticien au CERN (Laboratoire européen pour la physique des particules) de Genève en 1992. Aujourd'hui, de nouveaux explorateurs plus performants, souvent gratuits quand ils ne sont pas imposés par l'éditeur du système d'exploitation, et l'intégration de données hétérogènes mêlant textes, images, sons et animations contribuent au succès grandissant de cette interface. La notion de lien hypertexte est aussi une des clés de la réussite : une seule page HTML peut mettre à la portée d'un simple clic de souris une multitude de liens vers des ordinateurs dispersés sur l'ensemble de la planète ; et ceci de façon totalement automatique pour l'internaute. L'évolution du Web a permis de mettre à la disposition des utilisateurs des informations statiques, comme des catalogues, mais aussi des pages dynamiques qui sont créées à la volée, lors de la consultation. Ces possibilités de personnalisation du contenu en ligne sont sans doute un des éléments majeurs qui vont modeler l'avenir de ce réseau : deux internautes, dont les habitudes de consultation et les demandes précédentes sur un même serveur auront été automatiquement analysées, pourront se voir proposer respectivement deux pages avec des contenus totalement différents, malgré l'utilisation d'une même URL. Si on gagne effectivement en efficacité, on ne peut passer sous silence l'émergence de problèmes d'utilisation, par les gestionnaires des serveurs, de ces données personnelles pouvant représenter des atteintes à la vie privée de l'utilisateur.

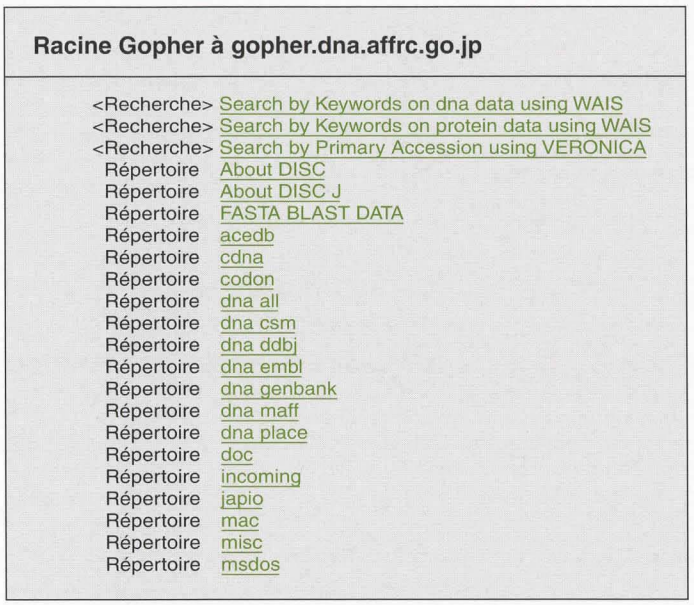

Figure 1. Un exemple d'interface de type « Gopher » avec accès à Veronica.

Figure 1. A Gopher-type user interface with Veronica access.

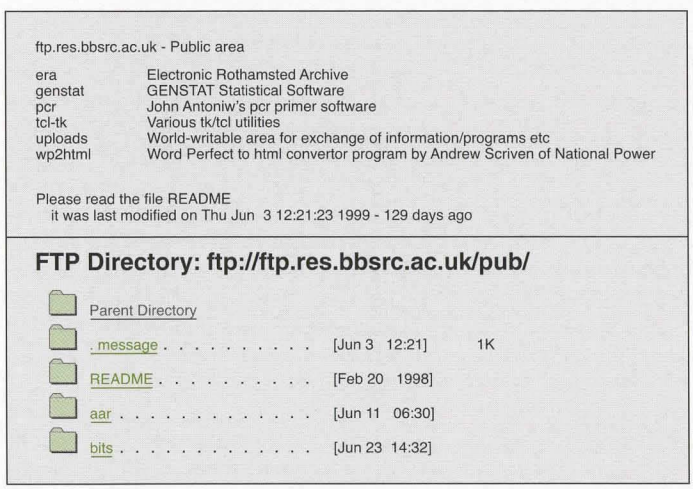

Figure 2. Un exemple d'interface de type « FTP ».

Figure 2. A FTP-type user interface.

permettent d'interroger plusieurs moteurs de recherche, d'éliminer les doublons entre les différents sites et de classer les références obtenues en fonction de leur pertinence. Les logiciels les plus évolués vont même jusqu'à vérifier la validité de l'URL *(Uniform Resource Locator)* référencée afin d'éliminer les références erronées ou celles pointant sur une page ayant disparu depuis leur première indexation par le moteur de recherche.

Afin d'évaluer l'importance des bases de données concernant les ressources génétiques sur Internet, nous avons utilisé un logiciel canadien, Copernic, pour explorer le Web, les sites FTP, Telnet, Gopher et les Newsgroups.

Copernic *(figure 5)* présente de nombreux avantages dont la gratuité de la version limitée. Celle-ci, comparée à la version commerciale, a cependant été suffisante pour mener à bien notre étude. Il permet d'interroger en parallèle 29 moteurs de recherche (18 anglophones et 11 francophones). Les requêtes se formulent très facilement, en incluant éventuellement les opérateurs AND, OR, NOT et NEAR, Copernic se chargeant d'adapter la demande au format des divers moteurs de recherche. Internet étant en perpétuelle évolution, une option du logiciel permet de mettre à jour automatiquement, par connexion sur le site de la société éditrice, les spécificités des moteurs de recherche interrogés, ce qui assure la mise à jour des caractéristiques de Copernic. Le résultat de l'interrogation des moteurs de recherche, limité à 1 000 références à la fois, peut être consulté sous forme de liste active : un clic de souris lance le chargement de la page web sélectionnée, qui peut aussi être exportée sous diverses formes de stockage. À ce niveau apparaît un autre avantage de Copernic, constitué par la qualité de la liste obtenue. Celle-ci est organisée par ordre de pertinence décroissante après que chaque référence a été validée.

Les mots clés que nous avons utilisés pour cette recherche ont été « germplasm AND database » pour les moteurs anglophones et « ressources génétiques AND base de données » pour les francophones (français et canadien). L'ensemble des titres, mais aussi le contenu des ressources disponibles sur Internet, a été exploré par les moteurs de recherches. En raison de la richesse du sujet traité et de la limite de 1 000 références récupérées à chaque requête, les interrogations ont été renouvelées, plusieurs fois par semaine, pendant 3 mois, entre mars et mai 1999. Malgré l'évolution rapide d'Internet et un certain hasard dans l'ordre d'apparition des 1 000 premières références trouvées, on peut penser qu'un aperçu quasiment exhaustif de la situation a été réalisé.

Traitement des données

Après vérification de l'existence des pages référencées par les moteurs de recherche, 2 069 URL uniques (dont 231 exclusivement avec les moteurs francophones) ont été analysées pour essayer de déterminer la meilleure stratégie de localisation des bases de données de ressources génétiques sur Internet.

Le traitement des références récupérées (URL) à chaque interrogation a été réalisé par un logiciel que nous avons spécialement conçu pour la présente étude. Ce logiciel réalise un pré-traitement de chaque référence afin d'éliminer les doublons obtenus d'une recherche à l'autre, en vérifiant le numéro IP du serveur. En effet, ce dernier peut posséder plusieurs identifications mais ne faire référence qu'à une seule et unique machine. La référence subit ensuite une analyse sémantique permettant d'identifier le pays abritant le serveur et l'organisation ou l'institution fournissant l'information. Cette référence est alors archivée, directement par le logiciel, dans une base locale au format Ms Access. La liste des moteurs de recherche ayant référencé cette URL est ensuite réactualisée. Afin de pouvoir déterminer le meilleur outil de recherche dans ce domaine, un tableau disjonctif a été constitué, repre-

Tableau 1

Principales listes de diffusion traitant des ressources génétiques

Nom	Sujet de la *mailing-list*	Adresse de souscription
BIO-AUTH	*Authority systems in taxonomy or systematic biology*	bio-auth-request@cmsa.berkeley.edu
CICHLID-L	*Systematics, ecology, behavior and conservation*	mailserv@nrm.se
CITRUS	*EGID-Citrus Network, Citrus related informations*	listserv@corse.inra.fr
EPD-L	*European Pollen Database List*	listproc@lists.colorado.edu
Genet	*Discussion group on plant genetic resources*	listproc@ucdavis.edu
Grain	*Genetic Resources Action International*	matthews@greengenes.cit.cornell.edu
MUSE	*MUSE Project*	listserv@cmsa.berkeley.edu
PEET-L	*PEET awardees funded by the National Science Foundation Partnerships for Enhancing Expertise in Taxonomy (PEET)*	peet-l@cmsa.berkeley.edu
TAXACOM	*Biological Systematics and Biocollections*	listserv@cmsa.berkeley.edu

Main mailing lists dealing with genetic resources

nant en lignes les différentes URL et en colonnes les différents moteurs de recherche. Ainsi, dans chaque cellule, la valeur 1/0 indique si cette URL est référencée ou non par le moteur de recherche. Ce tableau a été traité par ACP (Analyse en composantes principales) avec le logiciel Cstat. La représentation du plan factoriel principal fournit une image des principales stratégies d'indexation d'Internet par les moteurs de recherche. Ces stratégies d'indexation peuvent être interprétées comme des vues d'Internet données par chaque moteur de recherche.

Résultats et discussion

Quel moteur de recherche choisir ?

Un premier résultat marquant concerne la stratégie d'indexation des moteurs de recherche. Celle-ci est très différente selon le moteur utilisé, puisque le même nombre d'interrogations avec chacun d'entre eux, sur la même période, avec les mêmes mots clés, fournit de 38 à 675 références *(tableau 3)*. Cela ne permet d'identifier au mieux, avec un seul moteur de recherche, que 33 % des références totales.

Cette difficulté à trouver de l'information, malgré l'abondance des références existantes, met bien en évidence l'importance du choix du moteur pour réaliser la recherche. Si l'on considère les résultats de l'ACP *(figure 6)*, on observe que les deux premiers axes ne permettent d'identifier clairement que quatre moteurs. EuroSeek et Google semblent avoir une vision similaire du sujet recherché, alors que HotBot et Goto ont apparemment une vue très personnelle du Web. Parmi les six moteurs ayant indexé plus de 350 références chacun, AOL, Altavista et Excite ne semblent pas se distinguer des autres, au regard de l'analyse des autres plans factoriels de l'ACP (non représentés).

Un résultat surprenant est que plus des trois quarts des URL ne sont référencées que par un seul moteur de recherche. Seul un très petit nombre de serveurs sur les ressources génétiques est indiqué par plus de quatre moteurs, et aucun serveur n'est référencé plus de six fois parmi les différents moteurs possibles. Cela signifie que deux moteurs de recherche sur trois n'ont pas référencé une URL existante dans les index du troisième.

Il est donc extrêmement difficile de trouver une information donnée, malgré son existence sur le Web. Les moteurs sont nombreux mais très partiels dans leurs résultats, le référencement des pages étant très fortement lié au moteur utilisé. Cela met en évidence la nécessité d'utiliser un agent de recherche permettant d'interroger et de compiler les résultats de plusieurs moteurs [3].

Caractéristiques des bases de données sur les ressources génétiques

Ces caractéristiques ont été définies à partir de l'analyse de l'extension des URL. Cette dernière indique l'origine géographique du serveur pour les pays autres que les États-Unis (extension .fr pour la France, .br pour le Brésil…) et le type d'institution quand elle fait référence aux serveurs présents dans ce pays.

• Localisation géographique

Il apparaît très nettement *(figure 7* et *tableau 4)* que les serveurs américains occupent une place prépondérante (extensions des URL en .us, .edu, .gov principalement). Ils sont suivis de près par ceux des organisations internationales (extension .org). Ceux-ci représentent quand même une centaine de serveurs distincts dont la localisation géographique n'a pas pu être reportée sur la carte de la *figure 7*, car cette extension n'identifie que le type de structure fournissant les informations et non l'origine géographique. Les serveurs fournissant des informations sur les ressources génétiques sont répartis dans une quarantaine de pays, mais ceci ne préjuge en rien de la provenance géographique réelle des données consultables. Vouloir traiter de ce paramètre aurait nécessité de consulter individuellement l'ensemble des bases de données.

• Type d'institution

L'analyse des URL obtenues par les moteurs de recherche permet aussi d'avoir une vue générale sur les types d'institution qui mettent à disposition sur Internet des données concernant la biodiversité. Toujours en utilisant les extensions de l'URL, il apparaît clairement que les institutions nationales (gouvernements, organismes nationaux…) et les universités se partagent 85 % de la fourniture de l'information sur les ressources génétiques. Le reste (environ 15 % des références) est du ressort des organisations internationales (FAO,

Tableau 2

Principaux groupes de discussions traitant des ressources génétiques

Type de *newsgroup*	Nom du *newsgroup*
Divers *(ALTernative)*	alt.food.vegan alt.sustainable.agriculture
BIONET	bionet.maize bionet.microbiology bionet.plants bionet.plants.education bionet.software.acedb bionet.software.www bionet.virology
Informatique *(COMPuter)*	comp.ai.genetic comp.infosystems.www.anounce
Scientifique *(SCIentific)*	sci.agriculture.fruit sci.bio.microbiology sci.bio.phytopathology

Main newsgroups dealing with genetic resources

IPGRI…) et de quelques sites commerciaux (moins de 5 %). Pour ces derniers, l'accès est quelquefois soumis à un abonnement payant.

• Langue de consultation

Un aspect important, pour l'utilisateur de ces bases de données, est la langue utilisée pour la présentation de la page d'accès référencée par le moteur. Il apparaît clairement qu'il vaut mieux être anglophone pour parcourir le Web lorsqu'on est à la recherche d'informations sur les ressources génétiques. En effet, plus de 90 % des pages d'accès sont rédigées en anglais *(figure 10)*. On remarque que le français et l'allemand [4] se partagent les 8 % restants. Cette présence est, sans doute, le reflet de la longue tradition scientifique de ces deux pays dans le domaine de la taxonomie et la botanique. Cependant, la prépondérance de ces trois langues n'exclut pas la possibilité de pouvoir consulter certains serveurs dans d'autres langues. Ces derniers peuvent proposer sur la page principale une option donnant l'accès à cette même page rédigée en d'autres langues.

Contenu des bases de données

Une fois localisée grâce aux moteurs de recherche, le dictionnaire d'anglais éventuellement sous la main, la page tant recherchée est enfin à l'écran. Malheureusement, dans près de 75 % des cas, l'internaute n'aura pas accès physiquement à une base de données sur les ressources génétiques, mais à une page de liens hypertextes. En effet, plus des trois quarts des URL référencées par les moteurs de recherche sont constituées de liens permettant d'accéder à d'autres pages qui, dans 75 % des cas, pointent sur des pages de liens : la notion de toile d'araignée commence à prendre corps dans l'esprit de l'internaute... Malgré cela, ce concept de bibliothèque de liens est fort utile pour se constituer rapidement une sorte de bibliographie virtuelle sur le sujet.

Heureusement, certains serveurs proposent l'accès à des bases de données réellement présentes sur le site consulté. Dans ce cas, la majorité des bases de données concernent le règne végétal [5], mais souvent orientées vers une seule espèce [6]. Viennent ensuite les banques de données spécialisées sur le génome (carte génétique, marqueurs…) [7]. Les bases de données en ligne sur les animaux et sur les micro-organismes ne représentent que 25 % des cas.

Pour être chanceux, l'internaute s'intéressant aux ressources génétiques doit donc être anglophone, utilisateur des dernières technologies en matière de recherche d'information, bien organisé et chercher dans le domaine des ressources génétiques végétales. Malgré cela, il aura quand même encore quelques difficultés pour localiser son Eldorado.

Cependant, les évolutions actuelles, tant au niveau des outils de recherche de l'information que de la conception et l'utilisation des outils de la gestion de celle-ci, laissent entrevoir l'espoir d'une meilleure efficacité entre la formulation d'une requête et le résultat obtenu.

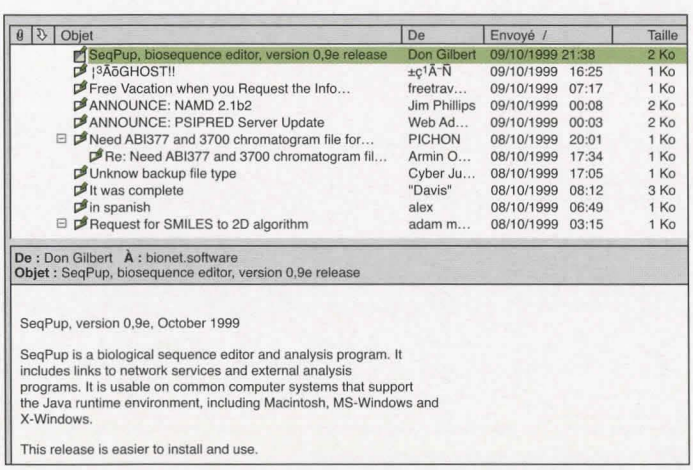

Figure 3. Une copie d'écran des messages postés dans le newsgroup « bionet.software ».

Figure 3. A screencopy from messages posted in « bionet.software » newsgroup.

Tendances et évolutions

À ce jour, il est difficile de faire une prévision réaliste, car la plupart des outils que nous utiliserons dans quelques années ne sont sans doute pas encore inventés. Malgré cela, certains verrous potentiels sont déjà identifiés et des pistes peuvent être dégagées pour tendre vers une homogénéisation à tous les niveaux. Le premier niveau est purement physique, c'est le format informatique utilisé pour le stockage des données. Un second aspect concerne une meilleure coordination au niveau international dans la conception et la définition des bases de données. Cela

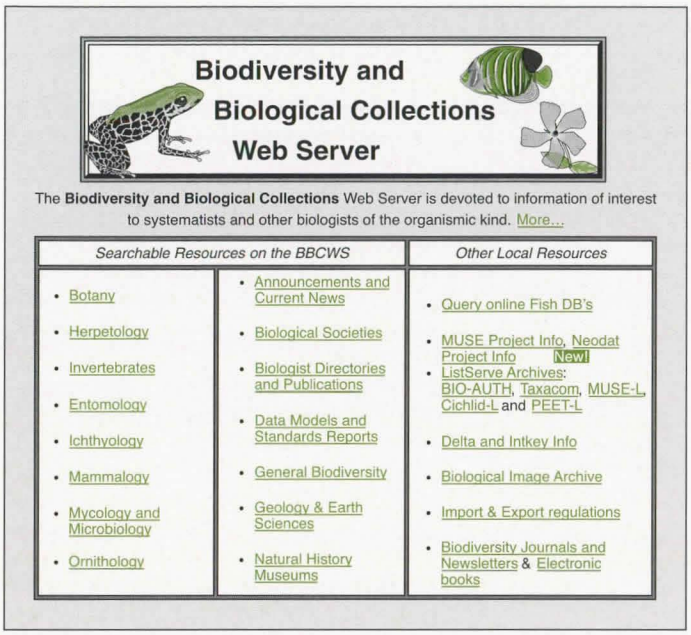

Figure 4. Un serveur WWW : la page d'accueil du serveur du BBC. Une liste d'hyperliens sur les principaux thèmes des ressources génétiques.

Figure 4. A Web server : a section of homepage for BBC server. A list of hypertext links on main genetic resources topics.

sur son serveur Web, selon un même format, les renseignements sur l'ensemble des collections gérées par le ministère de l'Agriculture des États-Unis (USDA), qu'elles soient végétales, animales, microbiennes. Si l'avantage d'une telle approche est clair au niveau conceptuel (indépendance du format et du sujet traité), il apparaît rapidement qu'elle ne peut conduire qu'à une description sommaire, étant donnée l'extrême diversité des caractéristiques biologiques des individus gérés par le système. En effet, la conception et la gestion d'un système unique visant à décrire de manière détaillée, par exemple, à la fois un agrume, une vache, un trypanosome... ne sont pas encore dans le domaine du réalisable. Il est à noter que la majorité des bases de données gérées par le Grin sont disponibles en téléchargement, accompagnées d'une interface rustique, mais efficace, permettant la consultation des données et leur exportation en local.

implique que les différents acteurs développent une meilleure collaboration au sein de groupes thématiques. Enfin, l'émergence d'outils de plus en plus autonomes, s'appuyant, entre autres, sur les dernières technologies d'intelligence artificielle, va permettre de simplifier les tâches de recherches aujourd'hui dévolues à l'internaute.

Formats communs

Si l'anglais semble être, de fait, la langue de consultation des serveurs traitant des ressources génétiques sur Internet, le codage informatique des données est loin de présenter la même uniformité. Parmi les nombreuses propositions pour faciliter les échanges dans ce domaine [8-15], deux d'entre elles semblent se développer plus particulièrement.

• DELTA

Le format DELTA *(DEscription Language for TAxonomy)* est une méthode, d'origine australienne, définie par un ensemble de règles permettant de coder les descriptions taxonomiques en vue de leur traitement par l'ordinateur [16]. Elle présente l'avantage d'une certaine adaptabilité pour une utilisation dans les différents règnes du monde vivant. Delta a déjà été adopté par de nombreux sites comme étant le format de base pour échanger les données. Les données codées sous Delta sont directement utilisables pour la production de descriptions en langage naturel et de clés de classification, interactives ou conventionnelles [17]. Pour faciliter le codage et simplifier l'utilisation de ce langage, de nombreuses interfaces sont disponibles sous divers systèmes d'exploitations (http://muse.bio.cornell.edu/delta/).

• GRIN

Le GRIN *(Germplasm Ressources Information Network)* [18] met à disposition

Groupes d'intérêt et « métadatabases »

Depuis quelques années, on observe l'émergence de groupes d'intérêt *(Special Interest Network)*, constitués d'un ensemble de personnes ou d'organisations se regroupant pour mettre en commun à la fois leurs compétences et leurs données. Les « métadatabases » représen-

Tableau 3

Pourcentage du total des URL trouvées par chaque moteur de recherches

Moteurs de recherches	Nombre d'URL	% référencé
Google	675	33
Excite	611	30
GotoCom	527	25
EuroSeek	438	21
Altavista	374	18
AOL	361	17
HotBot	301	15
PlanetSearch	233	11
WebCarwler	219	11
Snap	179	9
Yahoo	164	8
LookSmart	153	7
Magellan	151	7
InfoSeek	150	7
Lycos	139	7
MSN	113	5
Netscape	112	5
LycosTop	38	2

Percentage of total URLs found by each search engine

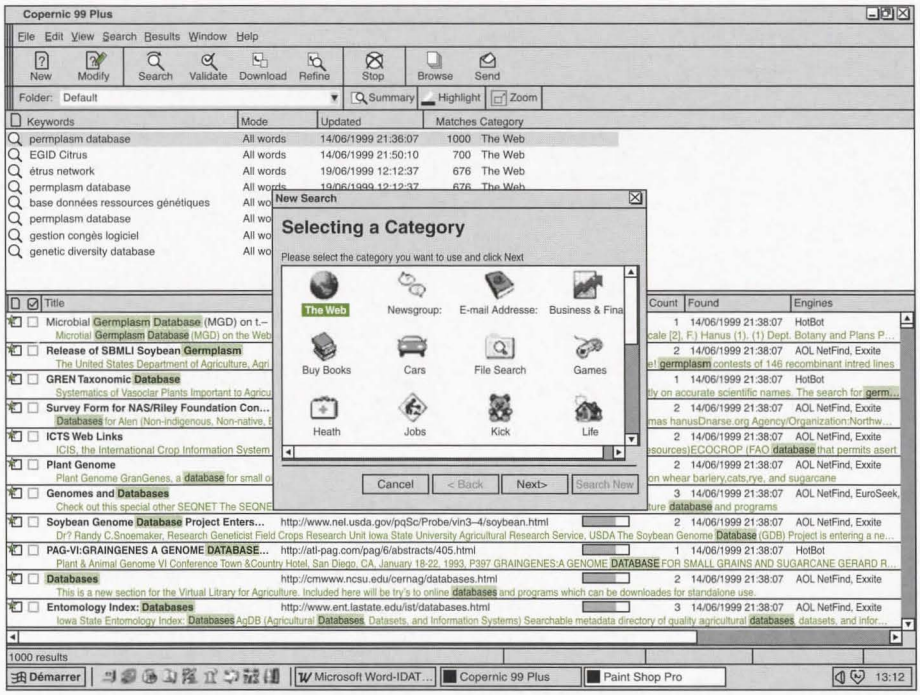

Figure 5. Copie d'écran de Copernic, agent de recherche.

Figure 5. Screen copy from Copernic, search agent.

tent actuellement un axe de développement important, commun à l'ensemble de ces groupes. C'est une interface offrant, à partir d'un point d'accès unique et d'une même requête, la possibilité d'interroger un ensemble de bases de données pouvant être localisées sur différents serveurs. Une liste d'URL, localisée sur un serveur, est l'expression la plus simple de ce que peut être ce type de produit. Par l'adaptation de l'interface en réponse à la demande de l'utilisateur, d'une part, et en incluant plusieurs sources d'information souvent hétérogènes, d'autre part, on obtient un outil extrêmement puissant pour extraire l'information *via* Internet. Par exemple, le projet GRID-INPE [19] fournit des références sur le thème de l'environnement, en incluant des données relevant des systèmes d'informations géographiques, du traitement d'images satellitales, en plus des données numériques et textuelles généralement disponibles dans ce type de base. Ce genre d'outil devrait pouvoir être utilisé pour diffuser efficacement l'information sur la biodiversité.

Les groupes d'intérêts, dont le statut juridique est très variable, assurent généralement, en plus de la mise à jour d'un serveur Web, la publication de lettres électroniques, l'animation de forum ou l'organisation de réunions de travail [20]. Nous avons choisi d'en présenter cinq qui illustrent la diversité de leurs domaines d'intervention et de leurs structures.

• TDWG

Le TDWG *(Taxonomic Databases Working Group)* existe depuis 1985 et représente une communauté d'intérêt, affiliée à l'*International Union of Biological Sciences* (IUBS). Il a pour objectif de standardiser les bases de données traitant de la taxonomie végétale et de promouvoir la collaboration entre les responsables de celles-ci. TDWG a étendu son champ d'action en incluant toutes les bases de données traitant de la taxonomie. La concertation au sein de ce groupe se fait essentiellement à l'occasion de réunions annuelles où sont discutés les aspects techniques relatifs aux bases de données, de même que la forme et le contenu des niveaux d'informations proposées. Ces réunions sont également l'occasion de partager les renseignements sur les développements en cours dans les bases de données sur la biodiversité. Ce groupe international, très actif, ne dispose pas d'un site Web fixe, il est généralement hébergé par l'institution qui organise la réunion annuelle.

Tableau 4

Nombre de serveurs fournissant des informations sur les ressources génétiques par pays

Pays	Nombre de serveurs
États-Unis	265
Organisations internationales	105
Localisations diverses	85
Royaume-Uni	38
Canada	26
Japon	23
Australie	21
Allemagne	18
Italie	14
Pays-Bas	12
France	10
Autres (32 pays)	190

Number of Web server providing genetic resources information by country

• IOPI

L'IOPI *(International Organization for Plant Information)* est une commission de l'*International Union of Biological Sciences* (IUBS) créée au début des années 90. Son but est de coordonner la création de bases de données sur les collections végétales. Son site Web (http://iopi.csu.edu.au/iopi/iopihome.html) regroupe un ensemble de liens assez exhaustif sur le sujet, avec une mention particulière sur l'utilisation du langage DELTA, un autre produit australien. Un effort particulier a été réalisé [21] au sein de ce groupe pour proposer un modèle conceptuel de données (MCD) applicable à l'ensemble des bases de données botaniques.

• Species 2000

De la même manière, mais de façon plus ambitieuse, Species 2000 a pour objectif de référencer toutes les espèces connues de plantes, animaux, champignons et microbes, pour favoriser les études de la biodiversité totale sur terre. Il fournit aussi une liste de points d'accès vers les autres bases de données, pour chaque groupe d'organismes vivants (http://www.sp2000.org/). L'utilisateur peut ainsi vérifier le nom scientifique, la classification taxonomique, de toute espèce connue *via* une interface qui fournit l'accès aux données extraites des collections, membres de ce projet. C'est un bon exemple de métadatabases de données associées à un groupe d'intérêt et

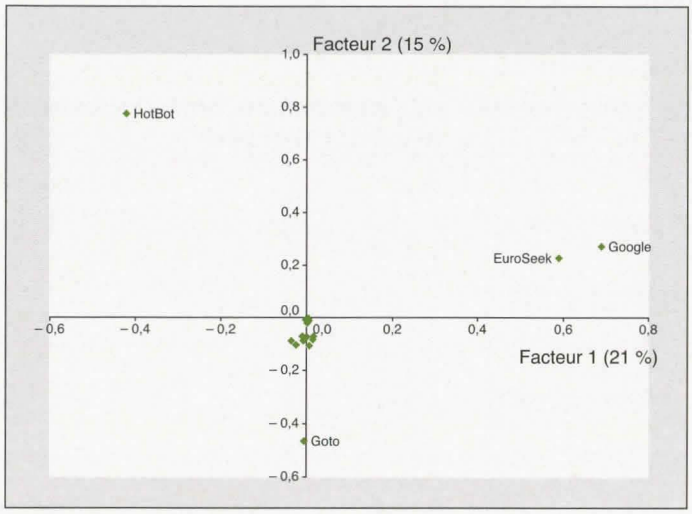

Figure 6. Variabilité des informations trouvées par les moteurs de recherches (ACP).

Figure 6. Variability of data indexed by search engines (ACP).

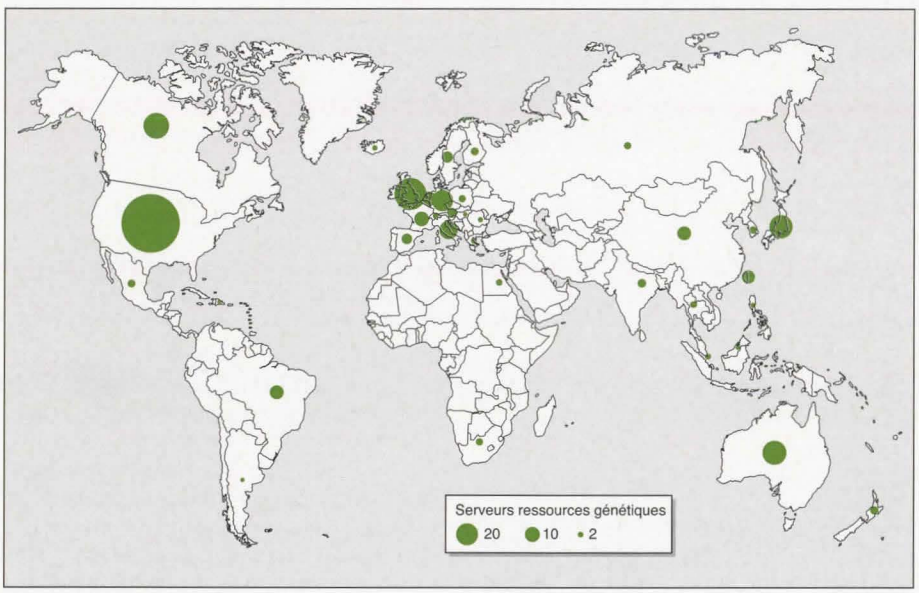

Figure 7. Localisation géographique des serveurs sur les ressources génétiques.

Figure 7. Geographic localization of genetic resources web server.

préfigure ce que pourrait être l'utilisation d'Internet dans ce domaine.

• ECP/GR

ECP/GR *(European Cooperative Program/Genetic Resources)* est un programme commun à la plupart des pays européens pour assurer la conservation à long terme et faciliter l'utilisation des ressources génétiques végétales en Europe. L'ECP/GR a été fondé en 1980 sur les recommandations du Programme de l'environnement des Nations Unies (UNEP), de la FAO [22] et d'EUCARPIA. ECP/GR est coordonné par l'IPGRI *(International Plant Genetic Resources Institute)*. Il vise à renforcer les liens entre les partenaires européens en favorisant, en particulier, le partage de la responsabilité de la conservation de celles-ci. Par exemple, la base de données *Prunus* européenne est maintenue par l'INRA (Institut national de la recherche agronomique) à Bordeaux. Cette base de données inclut les informations provenant de 26 pays européens sur les collections du genre *Prunus* et des espèces apparentées cultivées et sauvages. Sa mise en ligne sur Internet, au même titre que d'autres espèces du programme ECP/GR [23], est en cours de réalisation [24].

• BRG

Le Bureau des ressources génétiques (BRG) est un groupement scientifique français associant douze partenaires : six ministères et six organismes scientifiques. Le BRG a pour missions de créer un espace de dialogue pour les différents intervenants, de coordonner et d'expertiser la préservation des ressources génétiques animales, végétales et microbiennes en France.

De ce fait, il doit, entre autres, assurer la constitution et l'animation, au sein de réseaux, des ressources génétiques françaises. Ces ressources génétiques sont regroupées en Collections nationales ou internationales, lorsque la France, dans le cadre d'accords internationaux de coopération, a la responsabilité du maintien de celles-ci. La publication des coordonnées des responsables des collections, la consultation des catalogues, l'interrogation des bases de données sur le serveur Web du BRG (http://www.brg.prd.fr) sont en cours d'élaboration ; certaines collections sont déjà consultables en ligne.

Agents intelligents

La connaissance de l'existence d'une information et sa localisation restent aujourd'hui le principal problème lors d'une recherche sur Internet. Les moteurs de recherche actuels sont trop dépendants de leur mode d'indexation et de leur champ d'investigation. Une nouvelle génération d'outils, les agents intelligents, est en cours de développement par les grandes sociétés informatiques. L'intelligence de ces programmes est liée à leur capacité d'autonomie de déplacement sur le Net et à leur aptitude à analyser les données trouvées afin de proposer à l'utilisateur l'information la plus pertinente possible. Les applications prévues ne se limitent évidemment pas aux bases de données, sinon elles ne justifieraient pas l'engouement des grands groupes de l'informatique pour ce type de produit. Le principal objectif de ces entreprises est d'utiliser ce genre de logiciel pour le commerce électronique *(e-bussiness)*. Très prochainement, les agents intelligents seront capables de rechercher sur le Web le meilleur rapport qualité/prix pour un produit donné, et même d'en faire l'achat pour le compte de l'internaute... sous réserve que ce dernier prenne le risque de confier son numéro de carte bancaire à un petit pro-

Encadré 2

Glossaire

Les mots suivis d'un astérisque (*) sont définis par ailleurs.

Agent
Programme qui explore Internet et qui doit retrouver les informations recherchées. Les agents s'installent sur le disque dur ou fonctionnent directement sur un serveur Web. Dans ce cas, l'agent travaille hors connexion, sans consommer de ressources sur l'ordinateur local.

Forum de discussion
Voir *newsgroup**

FTP
File Transfert Protocol
Protocole de communication permettant d'échanger des fichiers entre ordinateurs.

Gopher
Système pour naviguer sur Internet avec des menus. Pratique pour accéder à des bases de connaissances. De moins en moins utilisé depuis l'avènement du Web.

HTML
HyperText Markup Language
Langage servant à décrire les pages Web et les documents hypertextes.

HTTP
HyperText Transport Protocol
Protocole de communication qui définit la façon dont les pages Web circulent de serveur en serveur.

IP
Internet Protocol
Identifiant unique d'un ordinateur connecté à Internet. Un numéro IP est constitué d'un ensemble de 4 nombres, séparés par des points. Chaque nombre représente un niveau du réseau : nom.domaine.organisation.ordinateur (Ex. : le numéro IP du serveur Web du Centre INRA de Corse est le 192.93.68.1)

Liste de diffusion
Forum de discussion* dont les messages circulent par e-mail. On s'abonne à une liste de diffusion comme à un magazine mais les abonnements sont généralement gratuits.

Mailing list
Voir Liste de diffusion*

Moteur de recherche
Site qui aide à retrouver d'autres sites. On l'interroge à l'aide de requêtes constituées de mots-clés ou en parcourant des index thématiques (sélection, répertoire).

News
Articles dans un *newsgroup**.

Newsgroup
Forum de discussion sur le réseau Usenet* traitant d'un thème particulier. Il en existe plusieurs dizaines de milliers. Le nom du *newsgroup* indique le thème de discussion : sci.agriculture.fruit est l'endroit où l'on parle de Science, et plus particulièrement des points relatifs à la culture des fruits. Tout le monde peut poster un e-mail dans un *newsgroup*, mais il est très mal vu de faire du hors sujet.

Newsletter
Bulletin d'information envoyé régulièrement par un site Web. Sur le site, on saisit son adresse e-mail dans un formulaire d'abonnement. La plupart des *newsletters* sont gratuites. Pour se désabonner, il suffit d'envoyer une simple commande par e-mail au serveur gérant la *newsletter*. Voir Liste de diffusion*.

Subscribe
S'abonner en anglais. On s'abonne à une liste de diffusion* à l'aide d'un message contenant la commande *« subscribe »*.

Telnet
Protocole Internet pour piloter à distance un ordinateur. Les utilitaires Telnet fonctionnent généralement en mode texte. Ils permettent, entre autres, d'accéder à des bases de données.

URL
Uniform Resource Locator
Chaque page Web dispose d'une URL ou adresse. Une URL commence par http:// (pages Web), ftp:// (fichiers à télécharger par FTP*), gopher:// (ressources Gopher*).

Usenet
Réseau qui rassemble les serveurs de *newsgroups**. Le plus souvent, ces serveurs ne sont pas directement connectés à Internet et nécessitent des logiciels spécialisés : les lecteurs de *news** *(News Reader)*.

Veronica
Very Easy Rodent-Oriented Netwide Index to Computerized Archives
Veronica est un type de moteur de recherche spécialisé exclusivement dans les recherches de type Gopher*.

www
World Wide Web
La « toile d'araignée mondiale », sous-ensemble de l'Internet. Le www regroupe des milliers de serveurs consultables par des pages HTML* et des liens hypertextes.

Encadré 3

URL

Ces URL sont valides au 1er octobre 1999.

7. Agent de recherche

- Copernic : http://www.copernic.com

8. Moteurs de recherche sur le Web

- AltaVista français : http://www.altavista.com/query?pg=q&what=web&kl=fr
- AltaVista : http://www.altavista.com
- AOL NetFind : http://www.aolnetfind.com/
- Carrefour : http://www.carrefour.net/
- Ecila : http://www.ecila.fr/
- EuroSeek : http://www.euroseek.net/
- Excite : http://www.excite.com/
- Fast Search : http://www.alltheweb.com/
- Google : http://www.google.com
- GoTo : http://www.goto.com/
- HotBot : http://www.hotbot.com/
- InfoSeek français : http://www.infoseek.com/Home?pg=Home.html&sv=FR
- InfoSeek : http://www.infoseek.com/
- La toile du Québec : http://www.toile.qc.ca/
- LookSmart : http://www.looksmart.com/
- Lycos France : http://www.lycos.fr
- Lycos Top 5 % : http://point.lycos.com/
- Lycos : http://www.lycos.com/
- Magellan : http://www.mckinley.com
- MSN Web Search : http://www.msn.com
- Netscape NetCenter : http://search.netscape.com/
- Nomade : http://www.nomade.fr/
- Planet Search : http://www.planetsearch.com/
- Snap : http://www.snap.com/
- Voilà : http://www.voila.fr
- WebCrawler : http://www.webcrawler.com/
- Yahoo France : http://www.yahoo.fr
- Yahoo : http://www.yahoo.com

9. Moteurs de recherche dans les *newsgroups*

- AltaVista Usenet : http://www.altavista.com/
- Deja : http://www.deja.com/
- Reference : http://www.reference.com
- NewsFerret : http://www.ferretsoft.com

10. Moteurs de recherche de *Mailing-list*

- Tile : http://tile.net/
- http://www.idf.net/mdr/glossaire.html

11. Les 13 liens incontournables pour commencer une recherche sur les bases de données de ressources génétiques

- http://www.agnic.nal.usda.gov/agdb/erdcalfr.htmlAgriculture-Related Information Systems, Databases, and Datasets
- http://muse.bio.cornell.edu/ Biodiversity and Biological Collections Web
- http://iopi.csu.edu.au/biological_information.html Biological information
- http://www.brg.prd.fr/ Bureau des Ressources Génétiques
- http://www.geocities.com/RainForest/Vines/8695/software.html Digital taxonomy software
- http://www.cgiar.org/ecpgr/platform/index.htm European Information Platform on Crop Genetic Resources
- http://www.fao.org/ Food and Agricultural Organisation
- http://www.ars-grin.gov/ Germplasm Resources Information Network (GRIN)
- http://125.ipgri.cgiar.org/ International Plant Genetic Resources Institute
- http://www.helsinki.fi//kmus/botsoft.html Internet directory for botany
- http://mbcr.bcm.tmc.edu/ Molecular Biology Computation ressource
- http://www.nbii.gov/biodiversity/index.html National Biological Information Infrastructure
- http://www.unl.edu/agnicpls/taxonom.html Taxonomy and Genetics

gramme se déplaçant de façon autonome sur Internet.

Conclusion

Cette étude avait pour objectif de faire un état des lieux des bases de données traitant des ressources génétiques sur Internet. Malgré l'abondance des informations sur ce thème, la difficulté d'aboutir de manière pertinente a été mise en évidence. En effet, l'utilisation d'un seul moteur de recherche ne permet d'obtenir qu'une faible proportion des références existantes. L'utilisation d'un agent de recherche, qui interroge et analyse les résultats obtenus par plusieurs moteurs de recherche, est indispensable pour contourner cette difficulté. Les bases de données recensées abordent principalement les domaines végétaux et génomiques, alors que celles afférentes au monde animal ou microbien sont peu représentées.

Si la langue utilisée majoritairement sur les serveurs est l'anglais, cette homogénéité ne se retrouve pas forcément au niveau du format des données mises à disposition. Il apparaît donc urgent d'organiser ces informations pour disposer d'outils fiables et performants, permettant à l'ensemble de la communauté d'accéder à la connaissance acquise. Cette organisation passe par la mise en place de règles à la fois de diffusion et de localisation, pour éviter les redondances et faciliter la recherche, ainsi que l'utilisation de formats communs. La tâche est immense, car elle doit être mise en œuvre au niveau mondial, avec des acteurs ayant des intérêts très différents, mais c'est à ce prix que Internet sera la « Corne d'abondance » que l'on espère trouver lorsque que l'on devient internaute ■

Références

1. Samier H, Sandoval V. *La recherche intelligente sur Internet, outils et méthodes.* Paris : Éditions Hermès, 1998 ; 155 p.

2. Chartron G. Recherche d'information sur Internet. In : Le Moal JC, Hidoine B, éds. *La recherche d'information sur les réseaux.* Paris : ADBS Éditions, 1996 : 43-101.

3. Dou C, Mannina B, Giraud E, Quoniam L. La méthodologie et la stratégie de recherche d'information à valeur ajoutée sur Internet. In : *L'information scientifique et technique et l'outil Internet.* Micro Bulletin thématique. Labège : CNRS, 1999 : 47-67.

4. Jimenez Krause D. Genetic resources on the Internet. In : *Das Deutsche Agrarinformationsnetz (DAINet).* Symposium, Juni 1995, Bonn. Bonn, Germany : Zentralstelle fur Agrardokumentation und-information, 1995 : 49-53.

5. Takeya M, Tomooka N. *The Illustrated Legume Genetic Resources Database on the World Wide Web.* National Institute of Agrobiological Resources 1997 ; 11 : 164 p.

6. Hummer KE, Strik BC. Strawberry genebank information on the worldwide web. *Acta Horticulturae* 1997 ; 439 : 49-53.

7. Kolak I, Satovic Z, Rukavina H. Plant gene banks in information and communication systems. *Sjemenarstvo* 1996 ; 13 : 253-60.

8. Anderson ML, Cartinhour SW. Internet resources for the biologist. In : *Biotechnology and plant genetic ressources : conservation and use.* Wallingford, UK : Cab International, 1997 : 281-300.

9. Beach JH, Ozminski SJ, Boufford DE. An Internet botanical specimen data server. *Taxon* 1993 ; 42 : 627-9.

10. Burley JR, Scott PR, Speedy AW. Biodiversity : the role of information technology in distributing information. In : *Biodiversity information : needs and options.* Proceedings of the International Workshop on Biodiversity Information, held in London, UK, July 1996. Wallingford, UK : Cab International, 1997 : 157-71.

11. Canhos VP, Manfio GP, Canhos DAL. Networks for distributing information. In : *Biodiversity information : needs and options.* Proceedings of the International Workshop on Biodiversity Information, held in London, UK, July 1996. Wallingford, UK : Cab International, 1997 : 147-56.

12. Green DG. Databasing diversity – a distributed, public-domain approach. *Taxon* 1994 ; 43 : 51-62.

13. Anonymous. Genetic Resources Action International (GRAIN) Genetic resources on the internet. *Biodiversidad : Cultivos y Culturas* 1996 ; 8 : 27-30.

14. Mincione A. *Internet and Intranet Technologies Applied to Germplasm Databases.* IMTAF, University of Reggio Calabria (RC), Italy, July 1997.

15. Schalk PH, Oosterbroek P. Interactive knowledge systems : meeting the demand for disseminating up-to-date biological information. *Biodiversity Letters* 1997 ; 3 : 119-23.

16. Dallwitz MJ. A general system for coding taxonomic descriptions. *Taxon* 1980 ; 9 : 41-6.

17. Aiken SG, Dallwitz MJ, McJannet CL, Consaul L. Diagnostic evidence from DELTA and clustering programs, and an INTKEY package for interactive, illustrated identification and information retrieval. *Can J Botany* 1997 ; 75 : 1527-55.

18. Wiersema JH. Taxonomic information on cultivated plants in the USDA/ARS Germplasm Ressources Information Network (GRIN). *Acta Horticulturae* 1995 ; 413 : 109-15.

19. Carling RC, Harrison J. Biodiversity information on the Internet : cornucopia or confusion. *Biodiversity Letters* 1997 ; 3 : 125-35.

20. Unger JM, Morin NR. The flora of North America project : a 21st-Century tool for managing plant information. *Am Biol Teacher* 1997 ; 59 : 338-43.

Summary

Genetic resource databases on Internet: access, present constraints and prospects

R. Cottin, E. Alfonsi, D. Agostini

In this paper, it has been done a review of databases on Internet as sources of information concerning the computerized genetic resources management in vegetal, animal, plant or microbial kingdom. A great number of URL specialized in the management of these genetic resources by means of databases has been obtained through a thematic questioning of search engines. An analysis has been done in order to identify organisms distributing this type of information on Internet, countries more active on this topic, and the genetic resources kind more frequently distributed on this support. In the following step, the relevance of tools in this kind of research is presented. Then, the possible evolutions are discussed, based on actual examples. A term glossary and a list of links necessary before beginning a research on Internet in the genetic resources domain complete this article.
Contrarily that one would be able to ingenuously believe, it is difficult to find a complete information on Internet (Table 3). *Only one search engine is insufficient to get information, search engines can have of strategies completely different* (Figure 6). *American servers are extensively dominant* (Table 4, Figure 7) *and the international institutes are less informative than it would seem at first sight. When one finally arrives to the source of information, one discover that these data are usually monospecific and that formats are not homogeneous.*
In conclusion, Internet is a computing system potentially powerful for searching information related to genetic resources. Nevertheless, a lot of work is needed to standardize the information in this databases and to improve the intelligence of servers to optimize answers.

Cahiers Agricultures 2000 ; 9 : 391-401.

Résumé

Cette étude recense les bases de données sur Internet, de même que les sources d'information concernant la gestion informatique des ressources génétiques, qu'elles soient d'origine animale, végétale ou microbienne. Grâce à une interrogation thématique par des moteurs de recherche, un grand nombre de références a été obtenu, donnant accès à des serveurs traitant de bases de données sur les ressources génétiques. Leur analyse a conduit à identifier les différents organismes diffusant ce genre d'information sur Internet, les pays les plus en pointe sur ce sujet et les types de ressources génétiques les plus fréquemment diffusées par ce support. De plus, la pertinence et l'efficacité des outils sont évaluées à travers cet exemple de recherche de l'information scientifique et technique. Les évolutions prévisibles dans ce domaine sont alors discutées en s'appuyant sur des exemples concrets. Un glossaire des termes utilisés et une liste de liens incontournables pour commencer une recherche sur Internet dans le domaine des ressources génétiques complètent cet article.

21. Berendsohn WG. A taxonomic information model for botanical databases : the IOPI model. *Taxon* 1997 ; 46 : 283-309.

22. Mangstl A, Judy JR, Ward FH. A new direction for FAO's information services – the World Agricultural Information Centre – WAICENT. *Agric Rural Develop* 1998 ; 5 : 32-6.

23. Boukema IW, van Hintum TJL. The European *Brassica* database. *Acta Horticulturae* 1998 ; 59 : 249-54.

24. Zanetto A, Formery B. International network on *Prunus* genetic resources : the European *Prunus* database. *Acta Horticulturae* 1998 ; 465 : 237-42.